# CONNEXIONS

## The Interoperability Report

April 1988

Volume 2, No. 4


*ConneXions - The Interoperability Report tracks current and emerging standards and technologies within the computer and communications industry.*

## In this issue:




ConneXions is published by Advanced Computing Environments, Inc., 480 San Antonio Road, Suite 100, Mountain View CA 94040, USA. Phone: 415-941-3399





ISSN 0894-5926


## From the Editor

TCP/IP is not, and never was, a US-only protocol suite. From the very beginning, researchers in Europe have been connected to the Arpanet, and much work has been done, particularly at University College London (UCL), on protocol development. This month, Jon Crowcroft of UCL gives an overview of some of this work. I am particularly grateful to UCL for providing me with access to the Arpanet while I was an undergraduate at the University of Newcastle upon Tyne from 1978 to 1983.

While we are on the subject of Europe and TCP/IP, there is a termendous interest in this protocol suite there now. Users are finding that TCP/IP provides them with interoperability while they wait for the promise of OSI. The latest proof of this interest is the CeBIT MultiNET demonstration which took place in Hannover, Germany in mid-March. *ConneXions* is working with attendees at this exhibition and we will bring you a report in an upcoming issue.

As described in our December 1987 issue, the Internet Activities Board (IAB) is the coordinating body for Internet research and development. The IAB consists of a number of Task Forces which work in specific areas. In a series of articles we will be looking at some of these Task Forces. First out is the Autonomous Networks Task Force which is described here by its chair, Deborah Estrin of the University of Southern California.

The Internet protocol suite supports a variety of network technologies with different MTU (Maximum Transmission Unit) sizes. This can lead to fragmentation of datagrams as they travel from one type of network to another. The effect of this fragmentation is explored in an article by Jeff Mogul of DECWRL on page 11.

Finally a reminder about our upcoming events: *The TCP/IP OSI/ISO Tutorials* (April 25 - 27), and the *TCP Performance Seminar* (May 9 - 10).

See you in Crystal City and Monterey!

# International Internetworking

by Jon Crowcroft Department of Computer Science, UCL

**Background**

The Department of Computer Science, University College London, (UCL) has a long history of Internetwork research. Since the early seventies, when UCL and the Royal Radar and Signals Establishment (RSRE) were the first European sites connected to the Arpanet via SATNET, and were also connected to the UK academic network (SERCNET now JANET), UCL has been working on protocol translation, and protocol performance. UCL has for some time been actively working on implementations of OSI protocols. Now that OSI convergence is a stated aim on both sides of the Atlantic, we are working with other organisations on management of OSI internets. This article outlines some of our current work in these areas.

**Stressful Internetworking**

The Department of Computer Science, University College London, is well situated for stressful internetworking. There is a 64kbps digital satellite service between the UK, Germany, Norway, Italy and the US, via SATNET, which has a minimum round trip delay of 2 seconds, and varies upwards of 20 seconds. In the past, it has also had high loss rates under load.

**Rich connectivity**

We have run an X.25 "tunnel" for some years, which gives us an alternative path to the US via IPSS and Telenet. This provides a maximum of 9.6kbps throughput. In the UK, we are connected to JANET (Joint Academic NETwork), a wide area X.25 network. Locally, we have a variety of LANs, gatewayed and bridged together, running TCP/IP, XNS, DECNET, X.25 on Ethernet (!!!), OSI protocols, and of course some other protocols.

We participate in the Admiral project which uses a high speed (2Mbps) experimental wide area network for distributed computing research. This network consists of five sites fully interconnected by a central switch, with the possibility of dynamic reconfiguration of inter-site bandwidth allocation.

**Many applications**

At a higher level in the protocol stack, we run a variety of applications (and associated session/presentation), with electronic mail, name services and directory services being the most rich in variety. We support several flavours of X.400 mail, as well as SMTP, UUCP and Grey Book. [The JNT "Colour Book Protocols" are a collection of standards developed for JANET, before the advent of ISO. JNT stands for Joint Network Team, the governing committee for JANET]. We use Domain Names, the JANET Name Registration Scheme (NRS), and THORN (THe Obviously Required Nameserver). We are implementing an X.500 directory service for RARE (European academic research network).

This complexity of different protocols and technologies has led to many interesting problems, and some interesting solutions which I will describe below.

**Protocol performance**

UCL has long been interested in making transport protocols work *efficiently* (in terms of loading the network), and *well* (in terms of performance for the user). Much work was done by Lloyd [10], Cole [1,2] and Edge [9] on adaption of retransmission timers and windows in transport protocols including TCP and ISO TP4.

Originally, SATNET was a microcosm of the current Arpanet behaviour. Recently, it has been observed that the characteristics of SATNET in terms of delay-bandwidth product match those of the high speed experimental networks (FDDI) that are emerging as technology of the future. Being able to adapt to loss, delay variation with load, congestion and so forth is an essential part of a good transport protocol implementation in a connectionless internet.

Before ISO, few protocol architectures had a well defined notion of Session or Presentation. Now that we have some working experience with these layers, UCL has been investigating their performance, and trying to see where they may be optimised.

Future high bandwidth networks may require special new approaches to exploit their capabilities. UCL has been involved in devising some new protocols for distributed systems, and has also been measuring the performance of David Clark's "rate based" protocol, NETBLT [7].

**TCP performance**

Some initial TCP implementations had fixed timeouts and simplistic adaptive windows. Usually these were optimal for a low delay, low error rate, uncongested LAN.

*Timers* : These simply failed when trying to run over SATNET, and so RSRE introduced the simplest possible feedback mechanism for adaptive retransmission timers, using linear feedback from the mean measured round trip delay. This mechanism fails to take account of the errors in one's estimation method or the underlying technological mean delay, and so several people (Edge, Mills, Jacobson et al) suggested that the mean + variance on round trip time should be used as a better estimator.

*Windows*: Many early TCP implementations use the receiver's advertised window, with the (often inaccurately) estimated round trip time to fill a pipe between the transmitter and the end of a receiver's buffers. The problem here is that the receiver closing the window is certainly a flow control mechanism. However, the receiver cannot tell the transmitter anything about network conditions, and so cannot use the same mechanism for congestion control. Loss of acknowledgements is used to back the transmitter off. However, many TCPs simply compound network congestion by retransmitting the entire window repeating an action that was a probable cause of loss in the first place.

**Slow-Start**

Recent developments (Jacobson/Karn/Partridge et al) [4] suggest a "slow start" to opening the window after loss, gathering speed with each acknowledgement until some threshold level of the previous problem window size. Then a linear increase (to try and find the optimum "congestion window" without oscillating window size around the right value). Current measurements over SATNET, either alone, or further on into Arpanet, show that these combinations have vastly improved TCP performance, from previous throughputs of 3-4 kbps, up now to 12 kbps. The real win is the number of unnecessary retransmissions, which has dropped effectively to zero. Future work includes the addition of a possible selective acknowledgement/retransmission option to TCP to deal with SATNET's unfortunate random losses.

*continued on next page*

## International Internetworking *(continued)*

**Session and Presentation performance**

UCL has been a close collaborator with Marshall Rose on the ISODE (ISO Development Environment), acting as a test site and UK distribution agent. We are using the Remote Operations Service, with a full ISO stack over X.25 (various flavours) as the basic building block for directory services.

Initial benchmarking showed that a Remote Operation to an actual application was around an order of magnitude slower than hand crafted raw use of transport alone. Initial profiling suggests that it is not an inherent feature of OSI upper layers that they are slow. Careful use of minimal copying and scheduling means that a single remote operation (excluding server processing) should take perhaps 3 times the raw transport time to complete. The culprit in current measurements is suspected to be application mis-use of the interface. (The important fact learned here was that "layer" is *not* equivalent to "process" and certainly does not entail value copying). References to buffers should be all that's required right as far as the operating system boundary.

Contrary to popular belief, the presentation syntax, ASN.1, is neither as space hungry, nor as expensive in processing as expected. When used properly, and with Macros, it is also a powerful specification tool.

**Transaction and blast-oriented protocols**

UCL is involved in new protocol work by being represented at the End to End Task Force. A particular area of interest has been the development of transaction protocols. A number of applications (e.g. Domain Servers, Routing Protocols) require sequences of reliably delivered messages - sometimes associated as requests and responses. The messages are sometimes large, as in booting systems, but are often small. There is a widespread requirement for a transport protocol to efficiently support these, requiring only two packet exchanges in the local network, small exchange case, while adapting to the wide area, large exchange case in a similar manner to TCP.

**NETBLT and VMTP**

UCL developed an experimental transport protocol (under BSD UNIX) to support such type of exchanges, which used some of the ideas from Birrel and Nelson [5] and some from Dave Clark's NETBLT protocol. NETBLT was developed as a new way of achieving high throughput with a novel flow control mechanism based on packet rates. Cheriton's VMTP has emerged as a strong Internet candidate for such a protocol, and it incorporates some of the other desirable features of modern protocols, including support for reliable multicast. UCL is currently engaged in measurements of NETBLT's ability to handle SATNET behaviour.

**Protocol translation and gateways**

Up to five years ago, all internal hosts at UCL were on Cambridge Rings, and ran the Ring protocols. To provide access to Public Data nets, JANET, and the Internet, we devised a unified transport service, that ran over the transport protocol, and built a gateway that front-ended UCL onto all the outside networks. More recently, we are confronted with the existence of ISO applications running over TCP/IP networks, X.25 networks, and pure ISO networks. Now that the predominant product for LANs is TCP/IP based, we are faced with interconnecting hosts over PDNs which do not support the same stack.

As ISO protocols become more widespread, gateways will support both DoD and ISO CLNS datagrams. This may have consequences for routing, both for intermediate systems and hosts. Of course, the European preferred selection of ISO stacks uses the connection oriented network service, because of the success and predominance of PDNs in Europe, and the high international availability of X.25 services. Network level connection of X.25 and ISO CLNS is of course impossible, and other solutions have been devised. In the next sections, I outline some of the past and current work in these areas.

**Unified Transport Protocol translation**

Inter-Process Clean and Simple (IPCS) was devised as a common transport substrate to support TCP, Yellow Book (a JNT protocol) and BSP (Cambridge Ring transport). By implementing the superset of functions of these protocols, together with an "intelligent" partial source route addressing scheme, IPCS enabled all UCL hosts to access all hosts on JANET, the Internet, and locally. The destination address was appended with enough information to identify a first hop translating gateway, together with some information as to what transport parameters (TSDU size, support for interrupt, and so on) were allowed for this route.

**OSI Transport Service gateways**

UCL has been using the experimental ISODE FTAM and its own X.400 over ISODE presentation and session. We run these over both X.25 and IP networks. It is possible to gateway between these stacks at the transport service, and UCL runs a service, switching FTAM at this level. One side of such a gateway runs TP0 service over Pure ISO connection (X.25) oriented networks. The other side treats TCP as a reliable network service, and runs TP0 over it. The application then runs through such a gateway. To automate relaying at this level then only requires a forwarding address mechanism.

**X.25 Tunnels and Bridges**

Two uses of X.25 PDNs have been in place for some time here. One was the X.25 Tunnel. The other a simple bridge. By mapping IP destinations into remote DTEs, and encapsulating IP in X.25 data packets, we were able to build an alternative route to SATNET to the US. Two difficult problems emerged:

1. Call management (when to create and destroy X.25 calls) became difficult when the round trip delays were high. IPSS charges are high, and there is a call setup charge as well as a per segment charge, and so optimising the lifetime timeout on a call is complex.

2. Advertising return path through the tunnel is problematic. The datagrams originated from the same UCL source as the ones going over SATNET, and so the natural return route from US hosts is via SATNET. Advertising the route when the tunnel opens is not quick enough, and costs money in any case. This problem has yet to be solved in general.

We are using LAN bridges that can operate between distant sites using any WAN line to connect the sites. Currently, leased lines, or ISDN can be used. However, private (PVC) X.25 network use is a viable alternative.

**Internetwork management**

Managing large networks is difficult. UCL is in a unique position of having to manage the interconnection of very diverse internets. Two basic building blocks are being developed to help in the areas of configuration, location and performance management.

*continued on next page*

## International Internetworking *(continued)*

To manage configurations of our systems, including host names and addresses, routing tables, mailbox location, relay service location and so on, we are building X.500 based directory services.

To manage performance, in terms of throughput, delay, availability etc., we are experimenting with automatic learning systems, based on principles from expert systems and AI technologies. Below, I outline some of this work.

### Directory Services

The CCITT X.500 recommendations outline how distributed directory services should be constructed and accessed. Currently we use Domain Nameservers, the Name Registration Scheme and the THORN DUA/DSA database, for managing most of these services. Work is ongoing to unify access to all this information, and this is seen as a key feature of protocol evolution.

The JNT have a firm plan to migrate to OSI over the next few years, and the plan includes the systematic control of changes. To help maintain connectivity, directory services will contain not only simple protocol services available at any site, but also will supply the location of protocol translators based on the context of the source of any query, as well as the destination. This will make an orderly and transparent changeover from the existing protocol stack to OSI, host by host, and site by site. UCL will apply similar techniques to support internetworking with other migrating networks.

### Learning Autonomous Systems

Many internets are now so complex that systematic analysis takes many experts a long time. Just one example of this is the time it has taken for transport protocol implementations to reach maturity. UCL is working on the use of learning systems to assist the tuning of performance of transport and network protocols.

A promising approach involves the application of genetic learning algorithms to this area. Genetic algorithms (proposed by Holland et al) [6] are particularly well suited to optimising discrete functions, where the search space for a solution is large, but solutions sparse. They also model populations of solutions well, and have been shown to result in "co-operating" solutions being found.

### Ecosystems

One can imagine a large internet as an ecosystem full of resources, and a collection of applications on hosts as a population of organisms trying to survive and use these resources. The hosts compete with each other for the resources, however, cooperating hosts have been shown to use the resources more effectively. In the presence of greedy hosts, a cooperative host loses out. The natural solution is for the environment (network/ecosystem) to react against greedy hosts by starving them of resources.

The implementation of this would be gateway algorithms like those suggested by Nagle [3] and Mills for selectively dropping packets from misbehaving hosts on a tit-for-tat basis. What UCL is developing is a general set of tools for generating such algorithms, in a protocol independent way.

## The future

The future is going to see more commonplace use of new facilities that are already emerging, including:

- Transaction Protocols
- Multicast over the Internet
- Extensive use of Directory Services
- Multimedia Mail and Real Time Conferencing

High Speed Networking (>Gigabits) is one hope for technology to support such applications. However, a good understanding of current systems is essential to avoid repeating past mistakes on future technology. This is where we hope UCL's main contributions lie.

*This work was partly carried out as part of the Autonomous Network Management project, supported under DARPA Contract No. N00014-86-0092. The author acknowledges this support. Any opinions expressed in this article are those of the author, and in no way represent an official view.*

## References

[1] R. Braden, R. Cole, P. Higginson, P. Lloyd "A Distributed Approach to the Interconnection of Computer Networks". Proc. ACM SIGCOMM 1983

[2] R. Cole, P. Higginson "Issues in Interconnecting Local and Wide Area Networks" Proc. Business Telecom Conference, Online, 1983

[3] J. Nagle "Congestion Control in IP/TCP Internetworks", Computer Communication Review, Vol. 14, No. 4, Oct. 1984

[4] C. Partridge, P. Karn "Improving Round-Trip Time Estimates in Reliable Transport Protocols", Proc. ACM SIGCOMM 1987

[5] A. Birrell, B.J. Nelson "Implementing Remote Procedure Calls", ACM Transactions on Computer Systems, Vol 2, No. 1, Feb. 1984

[6] J. H. Holland "Adaptation in Natural and Artificial Systems", University of Michigan Press, 1975.

[7] D. Clark "NETBLT: A Bulk Data Transfer Protocol", RFC 969

[8] D. Cheriton "VMTP: A Transport Protocol for the Next Generation of Communication Systems" Proc. ACM SIGCOMM 1986

[9] S.W. Edge "Connection Flow Control in a Wide Area Packet Switching Network", Ph.D. Thesis, TR 88, UCL, 1983

[10] P.J. Lloyd "End to End Protocol Performance over Linked Local Area Networks", Ph.D. Thesis, TR 123, UCL, 1986

**JONATHAN CROWCROFT** received his MA from Cambridge University in 1979, and his MSc from University College London in 1982. For the past 6 years, he has been working at UCL on a variety of networking and distributed systems projects including JANET, DARPA and Admiral (Alvey) related research. He is a member of the End-To-End Task Force, the Autonomous Systems Task Force and the Internet Engineering Task Force. Currently he is principal investigator for the DARPA funded Autonomous Systems project at UCL, and is active in the EEC funded RACE Nemesys High Speed Network Management project.

# The IAB Autonomous Networks Task Force

by Deborah Estrin, University of Southern California

## Introduction

The IAB Task Force on Autonomous Networks (ANTF) is concerned with communication across the boundaries of computer networks that are autonomously developed, owned, and operated. Research in autonomous networks (ANs) adresses two related problems: 1.Interconnection of two or more existing networks that are used and operated by separate administrations. 2. Decomposition of an existing network whose scale and scope make it difficult to modify and manage as a single, homogeneous network. Both situations raise issues related to technical and administrative heterogeneity. Currently we are focusing on issues of administrative heterogeneity namely, access control, authentication, and internet management.

## Access control across administrative boundaries

For example, we are considering the design of AN gateways and internet protocols that will respect and enforce administrative boundaries and related access control requirements. Can such mechanisms be added to existing internet protocols (i.e., IP), or is something other than IP needed for interconnection across ANs? Moreover, in order to implement meaningful access controls, an autonomous network must be able to authenticate communications from outside of its AN boundary. How can mutually suspicious autonomous networks, that share no mutual authority, establish a basis for authentication? Similarly, when an AN connects to the outside world the operating/environmental conditions of the gateways within the AN change. In particular, assumptions about the trustedness and reliability of routing and other network control information may no longer be valid.

Given the changes in environment and operating conditions that result from AN interconnection, the ANTF is re-evaluating the functional requirements for gateways, network routing protocols, and communication related applications.

## The Task Force

Like other IAB task forces, the ANTF is composed of researchers from academia and industry. However, this task force is newer than other well known task forces such as the IETF and is also somewhat more concerned with longer-term research issues of general interest than with nearer-term Internet engineering. In particular, the ANTF is not committed to working solely within the framework of IP and other existing Internet protocols. Part of its charter is to investigate and recommend areas in which the current Internet architecture is not adequate for a truly multi-administration environment.

## Current activities

In this relatively early stage of existence, the ANTF activities have focused on 1. AN interconnection requirements as examplified by case studies; and 2. the mechanisms needed to address these requirements, in particular access control, billing, and network routing mechanisms.

## Case studies

In addition to the existing Internet, we use several other cases of AN interconnection as points of departure for discussion. We have discussed the emerging US Inter-Agency Research Internet, the European research internet, private corporate network interconnection, and, of all things, The Telephone Network!

**Access control and billing**

Entities within a single autonomous network (e.g., within a single organization) typically trust one another more than do entities in distinct autonomous networks. Membership in a common administrative organization provides a basis for a priori trust, or routinized mechanisms for establishing trustworthiness (e.g., by referring to a mutual authority). Although the actual level of trust varies from one organization to another, many aspects of an organization's activities are based on this assumption of the ability to trust, while important aspects of inter-organization activities are based on the inability to trust a priori (e.g., contracts, reliance on legal system).

When autonomous organizations interconnect their networks they typically do *not* want to create one transparent internet out of their respective networks. Rather, they want to allow for controlled interconnection among a subset of entities within their respective ANs. Therefore, one requirement is to design gateways such that they can implement the desired access controls. Can such mechanisms be added to IP, or is something other than IP needed for interconnection across ANs? One concern is how to get the necessary level of granularity at boundary points in order to do desired access control.

**Enforcement**

Assuming *enforcement* of access controls is done at AN network boundary points, i.e., AN gateways, an internet comprised of many ANs becomes an even more difficult place in which to locate and travel to a desired destination. A routing agent now needs to know about access control policies in addition to connectivity in order to get a packet from one place to another. In other words, we are distinguishing between policy enforcement on the one hand and navigation around a policy defined network, on the other. Enforcement must be done in gateways (as well as some endpoints) because you can't rely on endpoints and there are network resources to be protected. Nevertheless, endpoints need policy/access control related assistance in navigating an internet that is defined by policy restrictions as well as connectivity restrictions. Moreover, end points may need assistance in establishing a "subscription" to follow a particular path, i.e., find out what authorization/keys are needed (e.g. what visas and currencies are needed to take a particular journey), they also want to have some idea of the cost associated with the journey.

**Policy server**

One way of approaching the problem is to have a model of the internet at the source in the form of a policy server and have the policy server figure out routes that the source will be permitted to use and the expense and authorizations required to do so (e.g. visas). Given such a server, we will need the hooks for the policy-sensitive layer to pass information to the network layer, as discussed below.

**Authentication**

Billing among ANs requires a means of identifying and authenticating on either an internet-wide or pairwise basis. Identification alone is not enough if there is no means to assure the authenticity of the claimed identification. We are currently investigating billing in the telephone network as a possible model for data internet billing mechanisms. As with access control, part of the problem is enforcement, and another part is giving the end system the tools with which to determine expected cost, i.e., enforcement alone is not enough.

*continued on next page*

## Autonomous Networks Task Force *(continued)*

**Network routing**

Existing network routing mechanisms assume and require more *trust, homogeneity,* and *globally-shared information* than typically is available in a multi-AN environment. In particular, there is a need for 1. policy based routing; and 2. control of routing information distribution.

There is general consensus in the Internet community that some type of policy-based routing capability is desired. However, we do not yet understand what fundamental policies should be enforceable. What should policy based routing be able to do? What are the basic policy information parameters that routing protocols should be able to recognize. Other TFs and working groups are also concerned with policy based routing requirements and we hope to contribute to the general discussion.

**Authenticating routing information**

The other side of the routing issue is what routing information to pass to and accept from other ANs. For example, should you advertise reachability to places that are not reachable due to access control/policy restrictions. This comes up for domain name servers in mail relays as well as for IP gateways. The tables are built up based upon network reachability while authorization may be based on higher level person name or organization affiliation. How to then add policy information to routing/reachability information? A separate issue is the need to authenticate the validity of reachability information that you receive. If on AN is skeptical about routing information from outside either because of its origin or its content, can it impose local constraints on routing information exchange and calculation and still avoid loops? Or must the validity of routing information be uniform across the collection of interconnected ANs?

**General issues**

In addition to access control, billing, and routing, some general issues come up in our discussions of ANs. For example, how much can a gateway do per packet? and how much traffic will an AN gateway have to handle? This is an important parameter to keep in mind when designing additional functions for AN gateways and protocols. A second general issue comes up in the case of breaking apart homogeneous networks such as internet or telephone network to allow for multiple sourcing and operation; namely, at what point do boundaries get in your way, and in what cases are they expendable or malleable?

**Future plans**

The current incarnation of the ANTF has met twice times, and will have its third meeting in May 1988. At that meeting we will attempt to settle on one or two projects for more detailed study. At the same time we will continue to survey and investigate examples of AN interconnections and problems that arise therein. We welcome your input in the form of examples, horror stories, or impediments which you may have encountered in interconnecting your organization's Autonomous Network to the outside world.

**DEDORAH ESTRIN** is an Assistant Professor in the Computer Science Department at The University of Southern California in Los Angeles. She received her Ph.D. in Electrical Engineering and Computer Science (1985) and her M.S. in Technology and Policy (1983) from MIT, and her B.S in Electrical Engineering and Computer Science (1980) from University of California at Berkeley. In 1987 she was chosen as a National Science Foundation, Presidential Young Investigator for her research in network interconnection and security. Her current research focuses on the technical and organizational issues related to the interconnection of computer networks across administrative boundaries.

# Fragmentation -- Pros and Cons

**by Jeffrey Mogul, Digital Equipment Corporation, Western Research Laboratory**

**What is fragmentation?**

One of the beauties of the Internet is that it is made up of networks with dramatically different properties; this is what allows us to access distant hosts in the same way that we access nearby ones. The most obvious differences in network properties have to do with bandwidth; the difference between a 1200 baud serial line and an Ethernet is about four orders of magnitude. This discrepancy will always be with us, since it is always possible to build LANs that are faster than Wide-Area Nets (WANs).

**MTU**

Most network technologies place a limit on the length of a single packet. In order to maintain reasonable delay, throughput, and error characteristics, this "Maximum Transmission Unit" (MTU) varies as a function of the bandwidth. On low-bandwidth networks, one must use a low MTU to avoid tying up the channel. One would prefer to use as large an MTU as possible, however, because the cost of per-packet processing is often the dominant cost of a network implementation, and so on high-bandwidth networks we see higher MTUs. For example, the MTU on the Ethernet is about 1500 bytes, whereas on the ARPA Packet Radio network the MTU is about 250 bytes.

This mismatch in MTUs presents a problem in an internet, since a host on a high-MTU network can generate packets destined for a host on a low-MTU network. When a gateway receives a packet that is larger than the MTU of the "next-hop" network, it must do something to preserve the contents of the packet without violating the MTU limit.

**Two kinds of fragmentation**

One solution to this problem is called *fragmentation*, where the gateway turns one large packet into several smaller packets for transmission on the low-MTU network. There are two kinds of fragmentation, differing mostly in how the fragments are *reassembled*:

- *intra-network fragmentation*:
  All the fragments of a packet are sent along the same link and are immediately reassembled at the next node (either the destination or the next-hop gateway). This could be done below (and transparent to) the IP protocol layer; in fact, the ARPAnet actually transmits up to 8 internal packets to transfer one MTU-sized datagram.

- *inter-network fragmentation*:
  The fragments of the packet may be sent along different links. Because there may be no single gateway that subsequently handles all of the fragments, reassembly must be done at the final destination host, and the fragments must be visible in the IP protocol layer (so that they can be forwarded by IP gateways).

*continued on next page*

## Fragmentation -- Pros and Cons *(continued)*

The specification of the IP protocol [2] allows both kinds of fragmentation. What this has meant in practice is that very few networks provide intra-network, or *transparent*, fragmentation, and inter-network fragmentation is the mechanism used within the Internet. All IP hosts must be able to reassemble fragments into packets; the algorithms for doing so are in RFC 791.

Fragmentation is a valuable mechanism because it allows us to use different MTUs in different parts of the Internet to maximize performance, while still allowing hosts to communicate without requiring them to know the MTUs of all the intervening networks. In effect, the gateways provide an automatic translation service when it is necessary to correct a mismatch in MTUs.

**Fragmentation can cause problems**

Although fragmentation is necessary to insulate hosts from the variation in MTUs, use of fragmentation can lead to poor performance or even total communication failure. This is particularly true of inter-network fragmentation, as used in the Internet, and has led to some serious problems in actual use. In [1] these problems are described in great detail. Here I will simply give two examples to illustrate the problem.

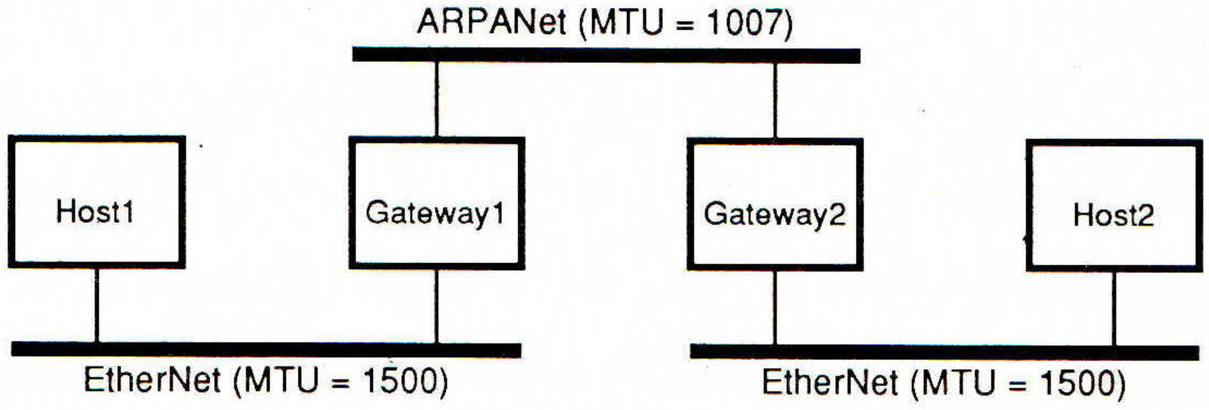


Figure 1: Situation where fragmentation may be necessary

First, suppose that Host1 is using TCP to communicate with Host2 (see Figure 1). Both hosts are on Ethernets, with MTUs of 1500 bytes, but the two Ethernets are connected only by the ARPAnet, with an MTU of 1007 bytes. Suppose further that Host1 wants to send 1024 bytes of data in each TCP segment. It sends its packets, which are received at Gateway1, and because a TCP segment containing 1024 bytes of data results in an IP datagram at least 1064 bytes long, Gateway1 must fragment the packet to fit it on the ARPAnet. The normal fragmentation algorithm creates a fragment of about 1000 bytes followed by one of about 100 bytes. The fragments are sent along the ARPAnet to Gateway2, which simply forwards them across the final Ethernet hop to Host2.

Suppose now that Host2 has an Ethernet interface that cannot receive two packets in quick succession; this is in fact true of a number of such devices. The second fragment may well arrive before the interface is ready, and it will be lost (figure 2 on the next page shows this happening). Since the datagram cannot be reassembled without all of its fragments, this means that the entire datagram is effectively lost. At some point, the TCP in Host1 will stop waiting for an acknowledgement for this lost segment, and will retransmit it ... leading, quite probably, to the same failure. Depending upon the probability of fragment loss, this data transfer will either proceed very slowly, or will not proceed at all.

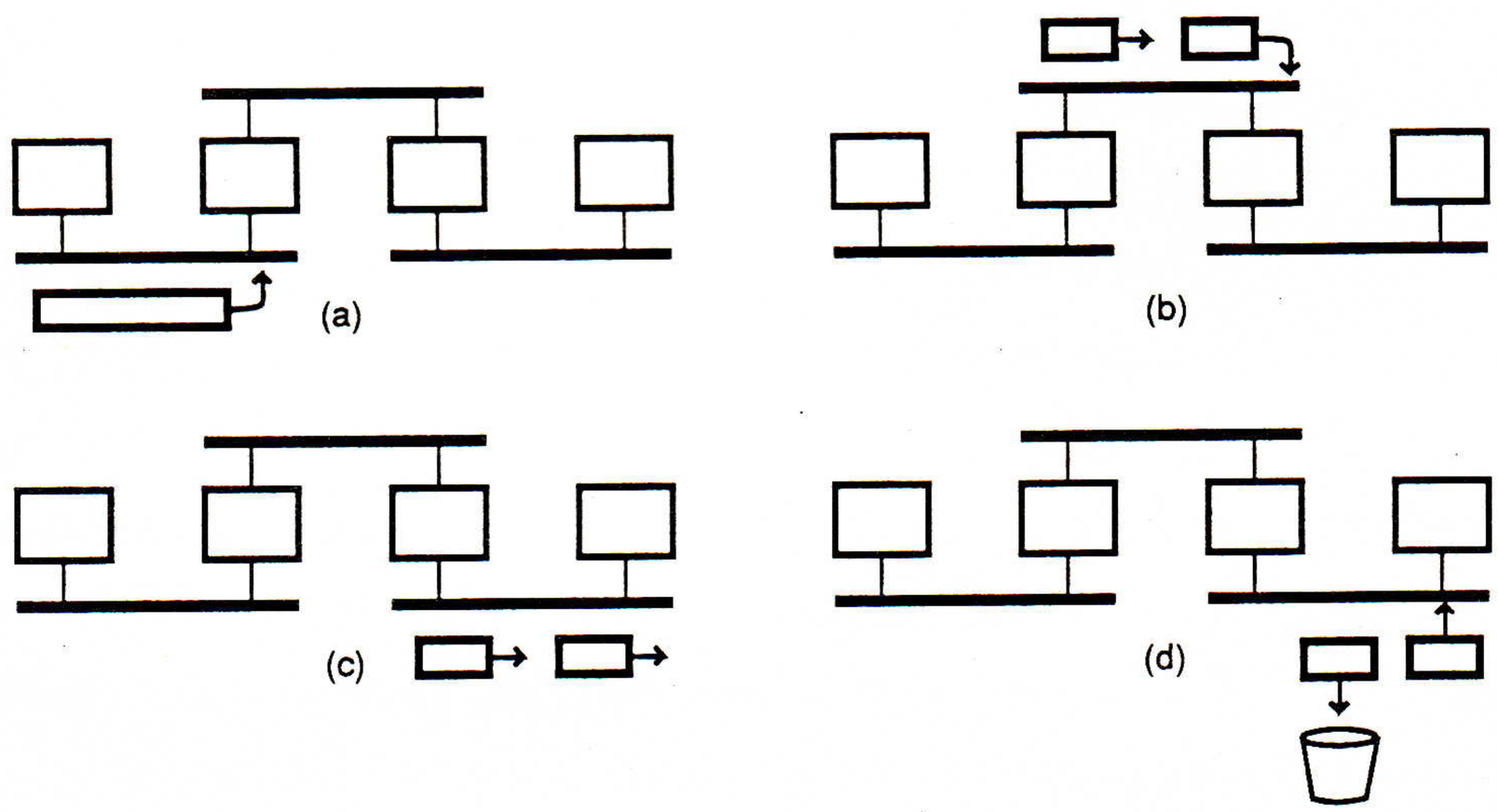


Figure 2: Example of deterministic fragment loss:
(a) packet sent
(b) packet fragmented by first gateway
(c) fragments sent on final hop
(d) second fragment dropped by destination

**NFS** Another example: the NFS protocol attempts to achieve good throughput by reducing the number of high-level protocol operations required to transfer a file. It does so by sending very large datagrams (8 Kbytes), which must be fragmented immediately by the sending host so as to fit on an Ethernet. This works fine if the destination host has a high-performance Ethernet interface and is on the same Ethernet. It breaks down badly when the packets must cross a gateway that might be even temporarily congested; such a gateway will drop some of the fragments. Repeated retransmissions by NFS do little good since the same thing may happen again; performance becomes poor or fails entirely.

Both of these examples illustrate two problems that conspire to cause trouble:

- *Deterministic fragment loss*: some packet loss mechanism that causes repeated loss of the same fragments.

- Fragments in IP are not individually acknowledged, so a lost fragment effectively means a lost datagram.

While the situation is actually a little more complicated than that, this description serves to lead us to several kinds of solutions.

**Avoiding fragmentation** One way to avoid the problems of fragmentation is to *avoid* fragmentation, by sending packets that will not need to be fragmented. This method may preclude maximum performance, but it improves reliability and avoids pathological performance problems.

*continued on next page*

## Fragmentation – Pros and Cons *(continued)*

To avoid sending packets that will be fragmented, the source host must know or estimate the minimum MTU on the path to the destination. In [1] we describe several ways to do this; the simplest solution is to send only small (576 byte) datagrams whenever the route involves a gateway. While this is a rather crude rule of thumb, it seems to work fairly well; this is the solution used in 4.3BSD and its many descendants.

A superficially similar method is used to avoid deterministic fragment loss for NFS. If an NFS client mounts a file system over a lossy path, it can specify that smaller transfer sizes should be used, thus reducing or eliminating fragmentation.

**Estimating MTU**

In [1] we also discuss various protocols for obtaining an accurate estimate of the actual minimum MTU on a path; since these protocols are not widely implemented, they are not yet a practical solution.

**Recovering from fragmentation problems**

Instead of avoiding fragmentation altogether, one can instead wait until problems arise and then reduce datagram sizes until fragmentation-related problems go away. The trick, in this case, is to detect that there is a fragmentation-related problem.

For example, a protocol implementation could infer from a high retransmission rate that deterministic fragmentation loss is occurring, and reduce the datagram size until retransmissions become infrequent. This works well for intentional fragmentation, such as with NFS, because it does not significantly increase the number of packets being sent. It is not a good idea for, say, TCP over a congested internet, since if the retransmissions are *not* due to deterministic fragment loss, use of smaller packets may actually increase congestion.

If we can actually detect fragment loss, rather than infer it, we can avoid that problem. Clearly the receiving host can detect deterministic fragment loss, since if fragments are lost then it will get "reassembly timeouts" (fail to reassemble a packet before the Time-To-Live value drops to zero). ICMP [3] already provides a mechanism for notifying the source host that this is happening; it is the "Time Exceeded" message (with a code of "fragment reassembly time exceeded"). If the source host receives one of these, it should reduce the size of datagrams it is sending to the destination host.

This mechanism can be used to estimate the actual minimum MTU of the path, and it requires no protocol changes. Unfortunately, few IP implementations actually send these ICMPs, and fewer do anything reasonable upon receiving them.

**Summary**

Improper use of fragmentation is a common cause of poor performance in the IP Internet. Contributors to this problem include:

- Interfaces and systems that can't handle back-to-back packets.
- Protocol designs based on intentional fragmentation, e.g. NFS.
- Failure to use the existing ICMP "Time Exceeded" mechanism to detect deterministic fragment loss.

Simple rules of thumb, such as limiting non-local datagrams to 576 bytes, resolve most of the problems without introducing too much inefficiency.

Future network designers should consider:

- Using intra-network (transparent) fragmentation whenever possible.
- Making it mandatory to handle error indications such as the ICMP "Time Exceeded"/"fragment reassembly time exceeded."
- Providing mechanisms in gateways and hosts to support determination of the actual minimum MTU along a path.

**References**

[1] Christopher A. Kent and Jeffrey C. Mogul. "Fragmentation Considered Harmful," Report 87/3, Digital Equipment Corporation Western Research Laboratory, November, 1987. Expanded version of paper presented at SIGCOMM '87.

[2] Jon Postel. "Internet Protocol," RFC 791, 1981.

[3] Jon Postel. "Internet Control Message Protocol," RFC 792, 1981.

**JEFFREY C. MOGUL** received an S.B. from M.I.T. in 1979, an M.S. from Stanford in 1980, and his PhD from the Stanford Computer Science Department in 1986. While at Stanford he produced the distribution of the Stanford implementation of PUP protocol software for UNIX. He is author or co-author of a number of RFCs, including those specifying Internet Standards for subnets, broadcasting, and RARP. Since 1986, he has been a researcher at the Digital Equipment Corporation Western Research Laboratory, working on network and operating systems issues for high-performance computer systems.

## Mark your calendars!

Our next conference is the *3rd TCP/IP Interoperability Conference* which will be held **September 26 - 30, 1988** at the Santa Clara Convention Center and the Doubletree Hotel in Santa Clara, California. It will feature the first *TCP/IP Interoperability Exhibition & Solutions Showcase* with the theme "TCP/IP - Today's Solution". TCP/IP technology, applications and products will be on display and vendors will show interoperability through cooperative demonstrations allowing attendees to learn about connectivity offerings in multi-vendor networks. We will bring you more information about this event in future issues of *ConneXions*.

**CONNEXIONS**
480 San Antonio Road
Suite 100
Mountain View, CA 94040

FIRST CLASS MAIL
U.S. POSTAGE
P A I D
SAN JOSE, CA
PERMIT NO. 1

**CONNEXIONS**

PUBLISHER Daniel C. Lynch

EDITOR Ole J. Jacobsen